Journal of Statistical Planning and Inference 187 (2017) 1–16

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Unbalanced ranked set sampling in cluster randomized studies

Xinlei Wang [a,*,1], Soohyun Ahn [b,1], Johan Lim [b,1]

[a] Department of Statistical Science, Southern Methodist University, 3225 Daniel Avenue, P O Box 750332, Dallas, TX 75275-0332, United States
[b] Department of Statistics, Seoul National University, Republic of Korea

ARTICLE INFO

Article history:
Received 18 March 2016
Received in revised form 2 February 2017
Accepted 13 February 2017
Available online 24 February 2017

Keywords:
Hierarchical linear models
Least squares
Missing data
Neyman allocation
Nonparametric inference
Order statistics
Ranking error
Relative efficiency
Treatment effect

ABSTRACT

We consider the use of unbalanced ranked set sampling (URSS) with cluster randomized designs (CRDs), and extend nonparametric estimators and testing methods, previously developed by Wang et al. (2016) for the use of balanced RSS (BRSS) with CRDs, to account for unbalanced stratified structures under different ranking schemes. We study the optimality, finite-sample and asymptotic properties of the URSS estimators, and numerically quantify and compare the relative efficiency of the URSS vs. BRSS estimators. We also study and compare the power of the URSS tests vs. their BRSS counterparts via simulation. Further, we investigate the application of the proposed methods to unbalanced data from BRSS-structured CRDs due to missing observations and illustrate it with an example using educational data. Finally, based on our results, we offer recommendations about when to use URSS/BRSS with CRDs.

1. Introduction

The effectiveness of new scientific methods is most convincingly established by randomized experiments, which often rely on simple random sampling (SRS) for randomization. Such experiments are often expensive and complicated to administer, and require recruitment of participants at various levels. Theory teaches that the most defensible scientific results come from using an entirely random process, for selection whenever possible, such as when selecting units at each level or assigning them to treatments. When sample sizes are small, especially, the vagaries of randomness can produce samples or treatment groups that seem unwise, in the sense that they do not represent the population adequately or are not similar enough to each other. It is well known that increased efficiency/power of inferential procedures based on SRS can be obtained by increasing the sample size. However, limitations on resources may prevent this approach. In such situations, ranked set sampling (RSS), which is a cost-effective method with a rich history (e.g., Chen et al., 2006; Wolfe, 2012, and references therein), can provide an alternative to SRS in randomized experiments.

Selection of a ranked set sample with a predetermined set size H begins by taking a simple random sample of size H from the population of interest. Then the H units are ranked by eye or some other cheap, but possibly imperfect, method that does not require measurement of the variable of interest. The unit ranked smallest among those sampled is measured

* Corresponding author.
E-mail address: swang@smu.edu (X. Wang).

[1] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

http://dx.doi.org/10.1016/j.jspi.2017.02.005
0378-3758/

and the rest discarded. A second simple random sample of size H is selected and ranked, and the second-smallest ranked unit is then measured and the remainder discarded. This process of taking a simple random sample, ranking, and measuring a single unit is continued through rank H (i.e., the largest rank), and a cycle is complete. These H ranks create H strata. More than one unit per rank may be acquired (say, n_h units for rank h, $h = 1, \ldots, H$). If m complete cycles are carried out, each rank stratum has the same number of replicates (i.e., $n_h \equiv m$) and the RSS design is called balanced; otherwise, the design is unbalanced. In this paper, we focus on the application of unbalanced RSS (URSS) in experiments with cluster randomized designs (CRDs), which typically involve clustered data (e.g., students within classrooms, patients within hospitals).

CRDs are widely used in educational, social and medical studies to assess treatment or intervention effects. Fig. 1(A) shows a typical example, where SRS is used to select schools (i.e., clusters) and then students (i.e., individuals) within each selected school. Let $Y_{k(ij)}$ denote the measured outcome of individual k in cluster j under treatment i ($i = 1$ for treatment and 0 for control), for $k = 1, \ldots, K_{j(i)}, j = 1, \ldots, J_i, i = 0, 1$, where $K_{j(i)}$ is the number of sampled individuals in cluster j under treatment i, and J_i is the number of clusters selected for treatment i. Note that, throughout this paper, we use k to index individuals, j to index clusters, and i to index groups; and we use "()" in subscripts to clearly indicate the nested structure, for example, $j(i)$ represents the jth cluster nested in treatment i. For correct analysis of data from cluster randomized studies, researchers often rely on hierarchical linear modeling. The model can be expressed by

$$Y_{k(ij)} = \mu + a_i + b_{j(i)} + r_{k(ij)}, \tag{1}$$

where μ is the mean score of the control group; a_i is the fixed effect of treatment i, with $a_0 \equiv 0$; $b_{j(i)}$ is the random effect of cluster $j(i)$; and $r_{k(ij)}$ is the random error, reflecting the effect of individual k in cluster $j(i)$ that has not been systematically accounted for by other terms in the model. The cluster effects $b_{j(i)}$'s are assumed to be identically distributed (i.i.d.), following some continuous distribution with mean $\mu_b = 0$ and finite variance σ_b^2; the errors $r_{k(ij)}$'s are assumed to be i.i.d. from some continuous distribution with mean $\mu_r = 0$ and finite variance σ_r^2. All $b_{j(i)}$'s and $r_{k(ij)}$'s are assumed to be independent. Under model (1), the treatment effect is given by $\Delta = \mu_1 - \mu_0 = a_1$, where $\mu_i = \mu + a_i$ is the mean score of the control/treatment group for $i = 0/1$.

Recently, Wang et al. (2016) have considered RSS-structured CRDs, where SRS is replaced by balanced RSS at different stages of the CRD (e.g., Fig. 1(B)–(D)), and developed nonparametric inferential procedures under a model-based framework, to achieve cost efficiency or better inference on estimating and testing the treatment effect Δ. Specifically, they have studied theoretical properties of the proposed RSS estimator under the hierarchical linear model (HLM) in (1), which has almost no distributional assumption. Further, they have formally quantified the magnitude of the improvement from using RSS over SRS, investigated the relationship between design parameters and relative efficiency (RE), and established connections with one-level RSS under completely balanced designs, as well as studying the impacts of clustering and imperfect ranking. All these have been done with balanced RSS (BRSS). So far, use of URSS with clustered data has not been addressed yet.

In the literature, a considerable amount of attention has been paid to URSS designs. Optimal allocation rules have been proposed to utilize some characteristics of the underlying distribution of data; and it has been shown that the performance of RSS can be greatly improved by appropriate unequal allocation (Kaur et al., 1997, 2000; Chen and Bai, 2000; Ozturk and Wolfe, 2000), especially for highly skewed data where BRSS does not work well in general. However, whether this is still the case when URSS is used with clustered data is an open question, which is worth investigating. Our paper is the first attempt to address the question and offer practical guidelines. When using RSS with CRDs, unequal allocation schemes can be adopted to recruit either clusters or individuals within (some) clusters, in which distributional information can be known from prior knowledge or past studies. One main purpose of this paper is to develop analytical methods to allow us to incorporate URSS into CRDs, and to determine when it is worth using URSS over BRSS or the opposite. Doing so would also offer the ability to handle practical situations when BRSS is used, but missing data often occur (at either level), leading to imbalance in different rank strata.

As in Wang et al. (2016), we assume judgment ranking so that no covariate information is needed for data collection or analysis. In the next section we present the notation we use for data from URSS-structured CRDs. In Section 3 we develop the URSS estimators of the treatment effect Δ under different ranking schemes, investigate their least square optimality and finite-sample properties, and evaluate and compare the relative efficiency of URSS vs. BRSS embedded within CRDs in various settings. In Section 4 we present the asymptotic properties of the URSS estimators, develop the asymptotic pivotal methods for testing Δ, and assess and compare the performance of URSS vs. BRSS testing procedures through simulation. To apply the proposed URSS methods to unbalanced data from BRSS-structured CRDs due to missing values, Section 5 discusses some theoretical justification and Section 6 presents a data example. Section 7 concludes the paper with a brief summary, presents guidelines about choosing between URSS and BRSS, and discusses potential future research. All technical proofs can be found in the Supplementary material (see Appendix A).

2. Notation for data from URSS-structured CRDs

There are three possible ranking schemes for a two-stage CRD: (i) ranking at the cluster level only (Fig. 1(B)); (ii) ranking at the individual level only (Fig. 1(C)); and (iii) ranking at both levels (Fig. 1(D)), where RSS, instead of SRS, is used to select clusters only, individuals only, and both, respectively.

For scheme (i), a separate ranked set sample of clusters is created for each treatment group. Let H_i denote the set size used in group i. For rank stratum h in treatment i, define the index set $\mathcal{J}_i(h) = \{j : \text{ cluster } j \text{ under treatment } i \text{ has rank } h\}$, where